Available online at www.sciencedirect.com

ScienceDirect

Procedia CIRP 33 (2015) 203 – 208

www.elsevier.com/locate/procedia

9th CIRP Conference on Intelligent Computation in Manufacturing Engineering - CIRP ICME '14

Statistical modeling of industrial process parameters

Francesco Aggogeri[a], Giulio Barbato[b], Gianfranco Genta[b,]*, Raffaello Levi[b]

[a]*DIMI, Università degli Studi di Brescia, Via Branze 38, 25123 Brescia, Italy*
[b]*DIGEP, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy*

* Corresponding author. Tel.: +390110907295; fax: +390110907299. *E-mail address:* gianfranco.genta@polito.it

Abstract

Identification of models of process parameters provides a way to clarify some hitherto unexplained patterns of deviation from design values, leading to enhanced opportunities of quality improvement. While most standard procedures are based upon normal distribution hypothesis, the latter sometimes is liable to fail to accommodate actual data even to a first approximation. Skew, bounded, multimodal data sets call for reasonably close description if meaningful inferences are to be drawn. Graphic representation may pose challenges, the aspect of grouped data being materially affected by a more or less arbitrary choice among several options. Issues in modeling are discussed in the light of an actual case, concerning a critical bore realization on an automotive component.

Selection and peer-review under responsibility of the International Scientific Committee of "9th CIRP ICME Conference"

Keywords: Process Modelling; Boring; Multimodal Distribution

1. Introduction

A stiff manufacturing schedule and tight specifications turned a bore finishing operation on a component into a foreman's nightmare, as performance of a complex manufacturing system was marred by scatter well beyond target capability, owing to a broad range of factors. A peculiar pattern of deviations from nominal diameter was observed, exhibiting inter alia a bimodal shape, as well as outliers affecting mainly one tail. In the quest for identification of main sources of trouble, statistical process modeling was resorted to, uncovering some problems concerning empirical distributions approximating those underlying data at hand.

Exploratory data analysis pointed to associations among process parameters and deviations from nominal diameter, leading to identification of steps susceptible to ensure process improvement. Machining of a cast iron component on a flexible manufacturing system was dictated by processing constraints, leading to problems linked to inherent system's complexity, compounded by a scheduling strategy dictated by tight requirements concerning production rate. Multiple fixtures were involved as well as different spindles and associated tooling, entailing additional sources of variation.

The case concerns a SME, tier one supplier of automotive powertrain components. Substantial investments were made in innovation technologies and human resources in order to increase the product portfolio, supplying special products for different requirements and applications. A surge in production volume with a downfall of increasing scrapped items suggested application of advanced statistical tools, in a drive to identify main factors affecting performances in current processes.

Quality issues surfaced concerning finish machining of a bore on a cast iron component, with tight specifications concerning diameter. The manufacturing system includes an interlinked set of CNC units, performing a range of machining operations including drilling, boring and grinding. A detailed process mapping was performed in order to identify, among the following list of potentially relevant factors, those requiring further investigation.

- Material: rough castings were provided by two different suppliers, chemical analysis being performed on incoming parts to check conformance with specifications before machining.
- Machine: the manufacturing system included an interlinked set of CNC machining centers earmarked for the specific

Selection and peer-review under responsibility of the International Scientific Committee of "9th CIRP ICME Conference"
doi:10.1016/j.procir.2015.06.037

operation at hand, whose parameters were mapped and investigated.
- Method: a set of measurements were collected to monitor production quality and yield, on a sample of pieces checked on a CMM at every shift. Bore finishing was performed either by boring or reaming, selected according to availability and set-up team criteria.
- Man: an operator loaded components on fixtures set on pallets shifted among machining units according to availability, with a dedicated set-up team on duty taking care of possible issues. Since production was carried out in two shifts, systematic differences were not unlikely.

Statistical analysis of production data indicated a rather poor fit to either Student or normal distribution; however the theoretical appeal of the latter, provided by the Central Limit Theorem (CLT), justified adoption of a mixture of normal distributions to provide an empirical model.

2. Modeling empirical data distribution

A set of over 600 parts machined in a pilot run exhibited a peculiar pattern of deviations from the reference value of the diameter of a critical bore, as shown by dot plot in Fig. 1.

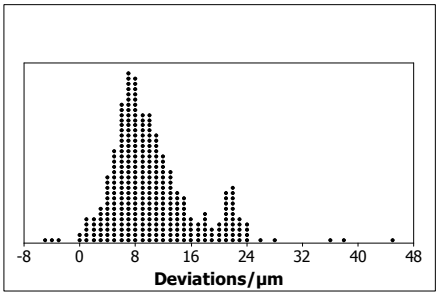

Fig. 1. Dot plot of deviations from reference value of diameter of a critical bore. Each dot may represent up to 2 observations.

A bimodal shape may be observed; furthermore, a few discrepancies appear on the left tail and some outliers may be identified on the right tail.

Outlier detection methods may be resorted to in order to identify discordant observations, a major shortcoming of most methods being the underlying hypothesis of normality, or even the requirement of knowing the underlying statistical distribution [1,2]. Given such knowledge, the problem of getting robust information from a reasonable number of data may be readily solved. When a few data only are available, difficulties are compounded by the fact that the main points of interest are on the tails, where data quality is inevitably poorer. Confidence or outlier identification intervals depend on probability concerning tails, and the difficulty of working in these regions appears evident. In fact some two centuries elapsed, since the groundbreaking work of Abraham de Moivre [3] on normal distribution, before a solution was provided to some practical tail problems by William Sealy Gosset [4] with his Student distribution.

Sound identification of statistical distribution on a purely empirical basis requiring a fairly large number of data, such an approach is ruled out in a number of instances. In the case at hand, the problem of outlier identification may not be approached in terms of the more common exclusion principles, as they are based on normal data distribution.

In the present work, an alternative method for outlier detection is proposed, based on an approximation of the experimental distribution with sound theoretical foundations. Some methods of exploratory data analysis are considered to model the empirical distribution.

At a preliminary level, histograms may offer a better representation of the empirical distribution of experimental data than dot plots, as bin width may be selected in order to highlight the most important aspects, a process entailing obviously individual appreciation. While there's no such thing as the "correct" bin width, some empirical rules provide a rough guidance, usually in terms of sample size n. Thus according to Sturges' rule [5] data range R is split into k equally spaced bins h wide, with

$$k \approx 1 + \log_2 n \tag{1}$$

A common rule in software packages, e.g. *Minitab*, requires:

$$k \approx n^{1/2} \tag{2}$$

In the case at hand, the number of classes is k=11 according to Sturges' rule and k=26 taking the square root of n, with corresponding bin widths of about 4.9 µm and 2.1 µm. Among a number of more or less similar rules, some take into account also measures of spread besides range (see e.g. [6,7]).

A good connection with the real situation can be given considering the concept of resolution, as described by VIM [8] in clause 4.14 :

> ***"resolution:***
> *smallest change in a quantity being measured that causes a perceptible change in the corresponding indication"*

Indeed resolution is a variability interval within which CLT works correctly, therefore it can be represented by a small normal distribution. This gives an indication justified by conceptual composition, even if its direct application is not easy: in fact resolution, as defined by VIM, depends on the measurement contest. The concept of reading resolution, also defined by VIM in clause 4.15, provides an easier approach:

> ***"resolution of a displaying device:***
> *smallest difference between displayed indications that can be meaningfully distinguished"*

Reading resolution is a well-defined, readily known characteristic of the measuring instrument concerned. As real variability is also affected by other factors, direct use of reading resolution as bin width would lead to an over-detailed description. A practical approach connecting such a readily

available information as reading resolution, with a reasonable description of the distribution, is provided by the method of kernel density estimation [9], a non-parametric method closely related to histograms offering however additional advantages, such as smoothness and continuity. Given a random sample $X_1,\ldots, X_n$ with a continuous, univariate density f, the kernel density estimator is

$$\hat{f}(x,h)=\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right) \quad (3)$$

with kernel K and bandwidth h [10], where kernel K may be a suitably general probability density function, typically unimodal and symmetric about zero, centered right over each data point. The influence of each data point is thus spread over its neighborhood, and the contribution from each point is summed in the overall estimate. Bandwidth h is a scale factor controlling how probability mass is spread around each point, besides smoothness - or roughness - of density estimate, over which it exhibits a strong influence.

Kernel density estimation, a numerical oriented approach, is implemented in specific software packages. In this work software *R* has been adopted, default kernel K being the normal probability density function, and default bandwidth calculated from the "oversmoothed bandwidth selector" [10]. In the case at hand, the default bandwidth ends up corresponding about with the experimental resolution, i.e. 1 µm.

Histograms and corresponding kernel density estimates respectively with a bin width and bandwidth chosen according Sturges' rule, square root rule and "oversmoothed bandwidth selector" are shown respectively in Figures 2, 3 and 4.

Sturges' rule substantially leads to overlooking the bimodal aspect of empirical data distribution, readily brought to light by less restrictive rules such as the square root one and the oversmoothed bandwidth selector. Smaller bin width and associated larger number of bins cater for closer description of shape as well as enhancing irregularities likely to be due to chance only, therefore a compromise is required.

Application of kernel produces a continuous representation of probability density that can be used in connection with basic concepts of traditional exclusion principles. In our case Chauvenet's criterion [11] may be readily applied, with an overall risk of excluding a sound value given approximately by $1/(2n)$, where n is sample size. Accordingly, given the cumulative distribution of kernel estimate, the tail bounds corresponding to the above defined risk may be readily estimated.

In order to reduce the influence of outliers on evaluation of boundaries, a trial-and-error procedure was adopted, by tentatively excluding suspect data and then calculating the corresponding kernel estimate on the remaining data, iterating as required. Accordingly, the first three data on the left tail (up to -3.1 µm) and the last five data on the right tail (over 27.8 µm) are tagged as outliers to be further investigated.

A bimodal distribution shape is confirmed. The wide use, and the very name of normal distribution, is due to the fact that central limit theorem is a description of what frequently happens: the joint action of a number of random effects yields a normal distribution. However also systematic differences are often present, leading to a description in terms of normal distributions, representing random components, offset by parts corresponding to systematic differences.

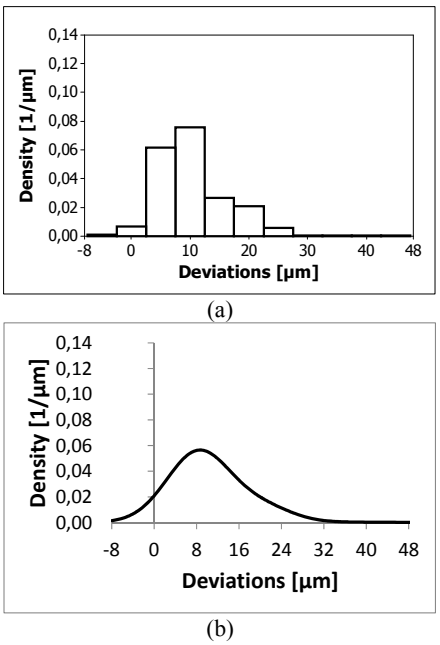

Fig. 2. Histogram (a) and corresponding kernel density estimate (b) with bin width and bandwidth respectively chosen according to Sturges' rule.

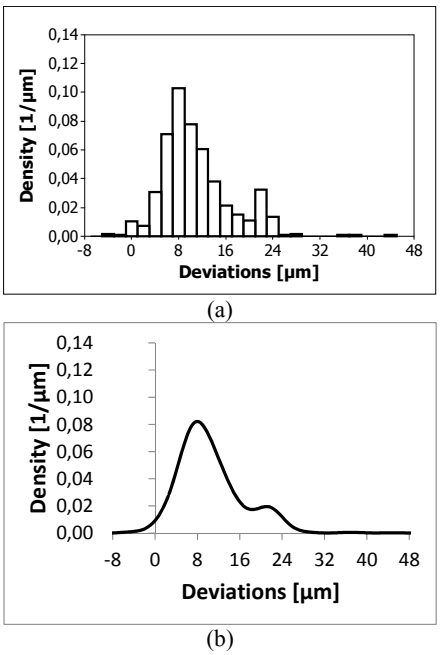

Fig. 3. Histogram (a) and corresponding kernel density estimate (b) with bin width and bandwidth respectively chosen according to square root rule.

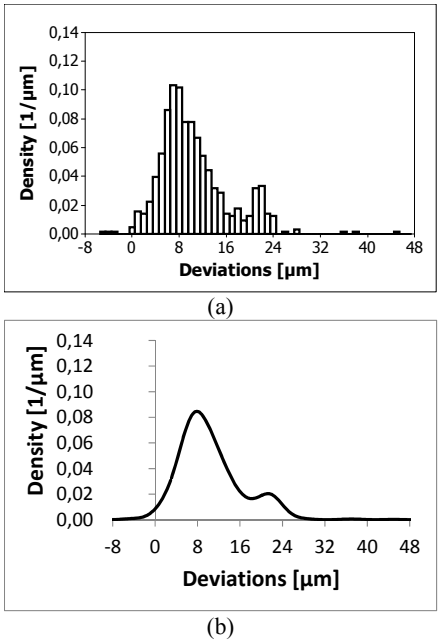

Fig. 4. Histogram (a) and corresponding kernel density estimate (b) with bin width and bandwidth respectively chosen according to oversmoothed bandwidth selector.

Such an approach may be described by a mixture of few populations normally distributed, explained in terms of relevant random and systematic effects. A mixture of two populations, present in data with percentages P_1 and P_2, normally distributed respectively with an averages μ_1 and μ_2 and standard deviations σ_1 and σ_2, whose estimates are given in Table 1, may justify the kernel distribution shape; but with a minor systematic shift in residuals, as shown in Fig. 5. A three component mixture, obtained considering also a third population present in a percentage P_3, having average μ_3 and standard deviation σ_3, estimated in Table 2, yields a closer fit as shown in Fig. 6; the improvement is however a minor one.

Table 1. Parameters of mixture of two normal distributions (corresponding χ^2 statistics is 3.24).

Distribution n.	*p*	*m*	*s*
1	86%	8.5	3.9
2	14%	20.8	2.7

Table 2. Parameters of mixture of three normal distributions (corresponding χ^2 statistics is 0.43).

Distribution n.	*p*	*m*	*s*
1	77%	7.9	3.6
2	11%	14.5	2.8
3	12%	21.5	2.4

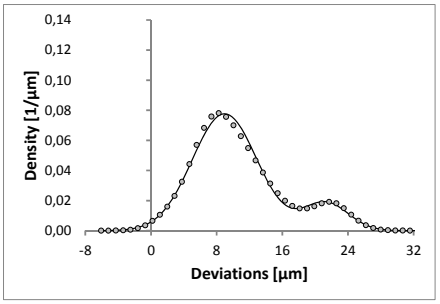

Fig. 5. Comparison between kernel pattern (continuous line) and a mixture of two normal distributions (dotted). A minor, albeit systematic, shift around both modes may be observed.

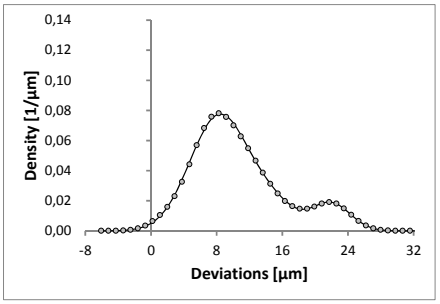

Fig. 6. Comparison between kernel pattern (continuous line) and a mixture of three normal distributions (dotted). The systematic shift present in Fig. 5 is almost eliminated.

The empirical estimates p_i, m_i and s_i yield indications concerning effects of the corresponding systematic factors. In the case at hand, identification was arrived at by comparing the kernel estimate with the mixture of normal distributions in terms of the relevant χ^2 statistics.

As evidenced by the mixture of two normal distributions, the process appears to be mainly affected by a factor centering data around 8 µm, exceeding by about 1.5 µm the mid-range of tolerance interval. However, additional factors produce a marked bias reflected by the upper part of the empirical bimodal distribution.

3. Discussion

The origins of the features of the empirical distribution were further investigated in terms of the main process factors, i.e. shift, setup team, supplier and tool, the latter two appearing to be dominant; a number of consideration were suggested by dot plots of Fig. 7 and mixtures of normal distributions of Fig. 8. The parameters of these mixtures are given in Table 3.

Bimodal shape appears to be associated mainly with boring, pointing out to systematic differences in tool setting criteria. Outliers on the left tail, appearing only when finishing by reaming, may be traced to measurement process, as observed elsewhere in industrial CMM work [12]. Right tail outliers, observed only on parts obtained from supplier 2, have as a likely root cause fixturing problems [13,14]. The effects of both supplier and tool are confirmed by two-way ANOVA

to be highly significant; they explain however only a fraction of total variation, the balance being due to other unidentified factors acting on machining system.

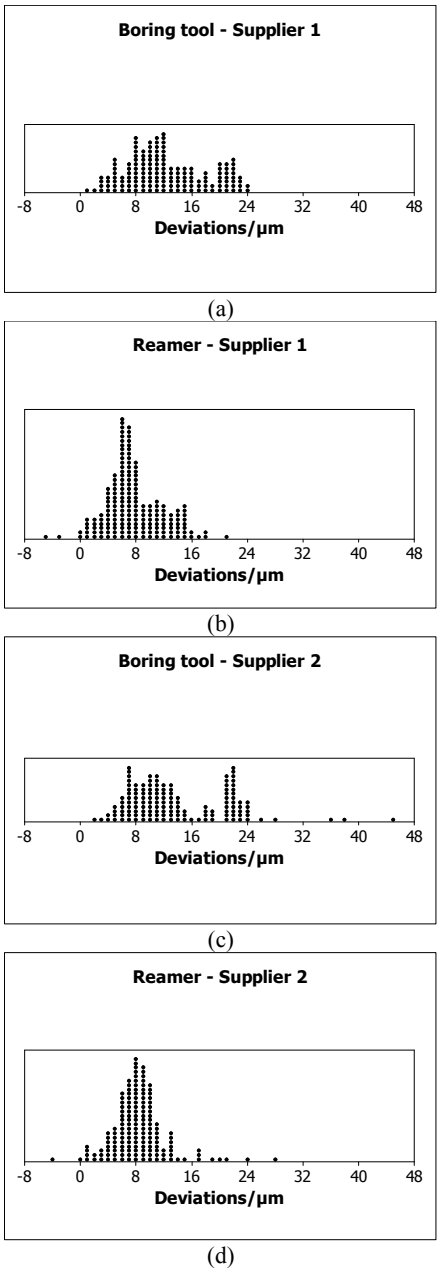

Fig. 7. Raw data subdivided according to tool and supplier.

Major departures from target process capability appear to be associated to boring, showing that the main part (Distribution n. 1) is biased toward averages around 10 µm, not acceptable according to tolerance limits specified. Furthermore, the other part (Distribution n. 2) is drastically biased toward averages around 21 µm, producing an even worse condition. This points out to substantial margins of improvement by taking such steps as enforcing uniformity of tool setting procedure, and of tool changing criteria.

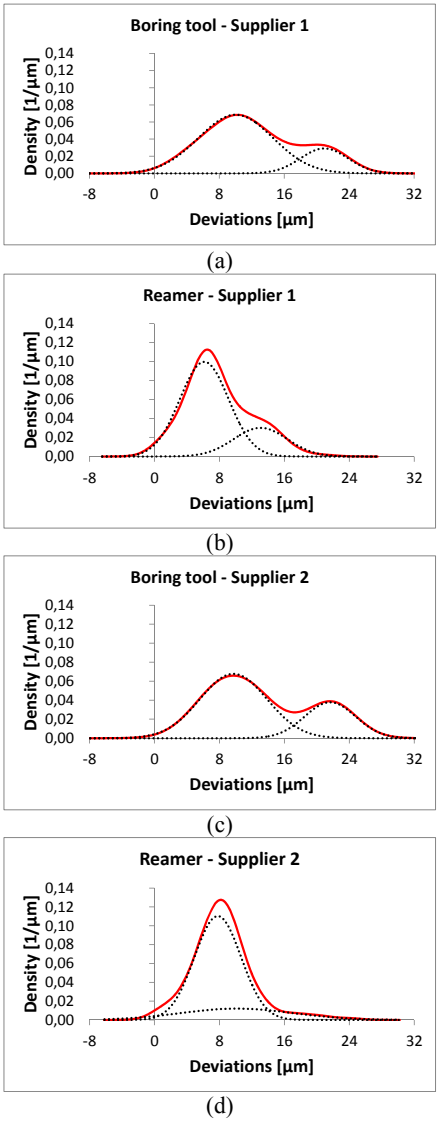

Fig. 8. Comparison between kernel pattern (continuous line) and the two normal distributions (dotted) which produce the mixture for data subdivided according to tool and supplier after outlier elimination.

With reamers, the main part (Distribution n. 1) is centered, with respect to tolerance limits, on about 6 µm for supplier 1, while is slightly biased around 8 µm for supplier 2.

Eventually, with boring tools Distribution n. 1 and n. 2 are significantly different for both suppliers. With reamers, the two distributions are significantly different only in case of supplier 1 (even if at a lower extent), while for supplier 2 the difference is hardly significant. In the latter case, Distribution n. 2 explains only an increased dispersion.

Empirical data modeling in terms of mixtures of normal distributions and kernel density estimation was thus shown to lead to detailed description of features pertaining to the data

set at hand, pinpointing technological aspects deserving further attention aimed at meeting exacting specifications in a cost-effective way.

Table 3. Parameters of mixtures of two normal distributions for data subdivided according to tool and supplier after outlier elimination.

Boring tool – Supplier 1			
Distribution n.	*p*	*m*	*s*
1	79%	10.0	4.6
2	21%	20.9	2.9

Reamer – Supplier 1			
Distribution n.	*p*	*m*	*s*
1	75%	6.1	3.0
2	25%	13.1	3.3

Boring tool – Supplier 2			
Distribution n.	*p*	*m*	*s*
1	70%	9.7	4.1
2	30%	21.6	3.1

Reamer – Supplier 2			
Distribution n.	*p*	*m*	*s*
1	79%	7.8	2.9
2	21%	10.2	6.9

References

[1] Barbato G, Barini EM, Genta G, Levi R. Features and performance of some outlier detection methods. Journal of Applied Statistics 2011; 38(10):2133-2149.
[2] Barnett V, Lewis T. Outliers in statistical data. 3rd ed. NewYork: Wiley; 1994.
[3] Student (Gosset WS). The probable error of a mean. Biometrika 1908; 6(1):1-25.
[4] De Moivre A. The doctrine of chances, or, a method of calculating the probability of events in play. London. 1718.
[5] Sturges H. The choice of a class-interval. Journal of the American Statistical Association 1926; 21:65-66.
[6] Freedman D, Diaconis P. On the histogram as a density estimator: L2 theory. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 1981; 57(4): 453–476.
[7] Scott DW. On optimal and data-based histograms. Biometrika 1979; 66(3):605–610.
[8] JCGM 200:2012. International vocabulary of metrology – Basic and general concepts and associated terms (VIM). 3rd ed. Sèvres. 2012.
[9] Silverman BW. Using kernel density estimates to investigate multimodality. Journal of the Royal Statistical Society - B 1981; 43: 97–99.
[10] Wand MP, Jones MC. Kernel Smoothing. London: Chapman & Hall; 1995.
[11] Chauvenet W. A manual of spherical and practical astronomy II. Philadelphia: Lippincott; 1863.
[12] Aggogeri F, Barbato G, Barini EM, Genta G, Levi R. Measurement uncertainty assessment of coordinate measuring machines by simulation and planned experimentation. CIRP - Journal of Manufacturing Science and Technology 2011; 4(1): 51-56.
[13] Müller P, Genta G, Barbato G, De Chiffre L, Levi R. Reaming process improvement and control: an application of statistical engineering. CIRP - Journal of Manufacturing Science and Technology 2012; 5(3):196–201.
[14] De Chiffre L, Tosello G, Píška M, Müller P. Investigation on capability of the reaming process using minimal quantity lubrication. CIRP - Journal of Manufacturing Science and Technology 2009; 2(1):47-54.